Package ‘KMDA’

April 1, 2015

Type Package
Title Kernel-Based Metabolite Differential Analysis
Version 1.0
Date 2015-03-26
Author Xiang Zhan and Debashis Ghosh
Maintainer Xiang Zhan <xiangzhan9@gmail.com>
Description Compute p-values of metabolite differential expression analysis using the kernel-based approach.
License GNU General Public License
Depends R (>= 2.10)
NeedsCompilation no
Repository CRAN
Date/Publication 2015-04-01 07:48:17

R topics documented:

KMDA-package .......................... 2
dkernel ................................. 2
dscore ................................. 3
hcc ..................................... 5
invlogit ................................ 6
mdist .................................. 7
pearson.group ........................ 8
skernel ............................... 9
spearman.group ..................... 10
sscore ............................... 11
tr .................................... 12

Index 14
Description

This package implements a kernel-based score test in metabolomic differential analysis. In order to capture the special natural of metabolomic data, two new kernel functions are designed in this package. One is a distance-based kernel and the other is a stratified kernel. This kernel approach also allows set-level analysis. It can be use to test whether a set of metabolites (or a metabolite pathway) are differentially expressed under two conditions.

Details

Package: KMDA
Type: Package
Version: 1.0
Date: 2015-03-26
License: GPL(>=2)

Functions: 
dkernel calculates the distance-based kernel.
skernel calculates the stratified kernel.
dsscore performs the distance-based kernel score test.
sscore performs the stratified kernel score test.
pearson.group performs the grouping of metabolites into metabolite-set based on Pearson correlation.
spearman.group performs the grouping of metabolites into metabolite-set based on Spearman correlation.

Author(s)

Xiang Zhan and Debashis Ghosh
Maintainer: Xiang Zhan <xiangzhan9@gmail.com>

References


Description

This function defines a distance-based kernel function.
Usage

dkernel(x, y, rho)

Arguments

x: a numerical scalar or vector of metabolomic measurements.
y: a numerical scalar or vector of metabolomic measurements.
rho: a positive real number, determining the smoothness of the kernel function.

Details

This function calculates a distance-based kernel function \( dkernel \) between two metabolomic measurements \( x \) and \( y \). It first calculates the distance between \( x \) and \( y \) (see function mdist for more details). Then the kernel function \( dkernel \) is calculated as

\[
dkernel(x, y) = \exp\left(-\frac{\text{mdist}(x, y)^2}{\rho}\right)
\]

Value

A positive real value.

References


See Also

mdist

Examples

\[
x \leftarrow \text{rnorm}(5)
y \leftarrow \text{rnorm}(5)
dkernel(x, y, 1)
\]

Description

This function test whether a metabolite-set is differentially expressed using a distance-based kernel score test.

Usage

dscore(x, y, lower, upper, m)
Arguments
- x: numeric measurements of metabolite abundance level.
- y: 0/1 response indicating whether a subject is a case group or a control group.
- lower: lower bound of the kernel parameter.
- upper: upper bound of the kernel parameter.
- m: number of grid points selected in the interval [lower, upper].

Details
Let $x$ be a $p \times n$ matrix, where each column is a subject, $y$ be a $n \times 1$ 0/1 vector indicating the group label. This function tests whether this $p$-metabolite set is differentially expressed between two groups (more details can be found in Zhan et al. (2015)). It works in the following way.

A score test can be applied when the kernel parameter $\rho$ is known. First, fit the null logistic model $\text{logit}(\text{pr}(y = 1)) = \beta_0$ to get estimate of $\beta_0$ as $\hat{\beta}_0$. Let $\hat{\mu}_0 = \text{invlogit}(\hat{\beta}_0)$. Second, The $n \times n$ kernel matrix is calculated as $K(\rho)_{ij} = k(x_i, x_j, \rho)$, where $x_i$ is $i$th column in $x$, $k(\cdot)$ is the distance kernel function $d_{kernel}$. Third, the test statistic $Q(\rho)$ is calculated as $Q(\rho) = (y - \hat{\mu}_0)^T K(\rho) (y - \hat{\mu}_0)$.

An standardized version $S(\rho)$ of $Q(\rho)$ can be calculated as $S(\rho) = |Q(\rho) - \mu_Q|/\sigma_Q$. More details can be found in Liu et al.(2008).

When the kernel parameter $\rho$ is not known. Suppose it takes values in $[\text{lower}, \text{upper}]$. Davies (1977) and Davies (1987) proposed a test based on the process $\{S(\rho), \rho \in [\text{lower}, \text{upper}]\}$. This test has rejection region of the form $\{\sup_{\text{lower} \leq \rho \leq \text{upper}} S(\rho) > c\}$. Using this test, an upper-bound for the p-value is given by:

$$\Phi(-M) + V \exp\left(\frac{1}{2} M^2\right)/\sqrt{8\pi},$$

where $\Phi(\cdot)$ is the cumulative distribution function of standard normal density, $M$ is the maximum of $S(\rho)$ over the range of $\rho$ and $V = |S(\rho_1) - S(\text{lower})| + |S(\rho_2) - S(\rho_1)| + \cdots + |S(\text{upper}) - S(\rho_m)|$ is the total variation of $S(\rho)$ over the interval $[\text{lower}, \text{upper}]$ and $\rho_1, \ldots, \rho_m$ are $m$ grid points in the interval $[\text{lower}, \text{upper}]$.

Value
A p-value indicating whether the metabolite-set is differentially expressed or not under two conditions/groups.

References
Davies, R. B. (1977) Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika, 64, 247-254.
Davies, R. B. (1987) Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika, 74, 33-43.

See Also

invlogit, dkernel

Examples

data(hcc)
x=hcc[1:3,3:U7]  ## this metabolite-set contains the first three metabolites in the hcc dataset.
y=c(rep(0,35),rep(1,20))
dscore(x,y,1,10,3)

Description

This dataset is a matrix containing measurements of metabolite abundance level.

Usage

data(hcc)

Format

A data matrix with 1388 rows and 57 columns. Each row is a metabolite. The columns are:
1st column: retention time;
2nd column: m/z (mass-to-charge) ratio;
3rd- 57th columns: abundance level measurements of metabolites from different subjects.

Details

This data are originally produced in Patterson et al. (2011). The size of this data matrix is 1388 × 57. Each row is a metabolite detected by some certain platforms. The first column is retention time, and the second column is the m/z ratio. Those two columns can be treated as identification of metabolites. The 3rd to 57th columns are measurements from 55 subjects. The column names indicate both the subject number and the group that subject comes from. 20 Subjects are from the Hepatocellular Carcinoma (HCC, n=20) group and 35 subjects are from the control group. Moreover, the control group can be divided into three subgroups. They are acute myelogenous leukemia (AML, n=22), healthy volunteers (H, n=6) and liver cirrhosis (LC, n=7). More details can be found in Patterson et al. (2011).

References

invlogit

Inverse Logit Function

Description

Given a numeric object return the inverse logit of the values. This function should not be called directly in this package, but be used by other functions like dscore and sscore.

Usage

invlogit(x)

Arguments

x A numerical value

Value

An object of the same type as x containing the inverse logits of the input values.

See Also

dscore, sscore

Examples

invlogit(0)
**mdist**  

*Metabolite Distance Metric*

**Description**

This function calculates a distance metric between two metabolomic measurements. These measurements can be either scalers or vectors.

**Usage**

```r
mdist(x, y)
```

**Arguments**

- `x`: a numerical scalar or vector of metabolomic measurements.
- `y`: a numerical scalar or vector of metabolomic measurements.

**Details**

If `x` and `y` are of different dimensions, function `mdist` returns a value of -1, which indicates the `mdist(x, y)` is not defined in this scenario. When `x` and `y` have the same dimension, suppose they have `p` components. If `p = 1`, then `x` or `y` is the abundance level measurement of a single metabolite, which is a non-negative real number. If `p > 1`, then `x` or `y` is measurements of a metabolite-set with multiple metabolites. In this case, let `x_i` be the `i`th component of `x`, which is non-negative and denotes the abundance level measurement of the `i`th metabolite in the metabolite-set. The distance between `x` and `y` is defined as:

\[
mdist(x, y) = \sqrt{\sum_i I[\delta_{x_i} \neq \delta_{y_i}] + \sum_i (x_i - y_i)^2},
\]

where \(\delta_{x_i} = 0\) if \(x_i = 0\), elsewise, \(\delta_{x_i} = 1\), and \(I[\cdot]\) is the indicator function.

**Value**

This function returns a non-negative value if `x` and `y` are of the same dimension. Otherwise it returns `-1`.

**References**


**Examples**

```r
x=c(0,1,2)  
y=c(1,0,3)  
z=c(0,1,2,3)  
mdist(x,y)  
mdist(x,z)
```
pearson.group

**Grouping Based on Pearson Correlation Coefficients**

**Description**

This function forms metabolite-sets based on pairwise Pearson correlation between different metabolites.

**Usage**

```
pearson.group(data, threshold)
```

**Arguments**

- `data`: a matrix with each row being a metabolite and each column being a sample.
- `threshold`: a threshold value for correlation coefficients.

**Details**

The input data is a matrix with each row denoting a metabolite. This function groups different rows of the data matrix together based on the Pearson correlation coefficients between two rows. It works in the following way.

First, each row in the data matrix is treated as a node. If the Pearson correlation coefficient between two nodes is larger than the threshold value, then an edge is added between this two nodes. Second, all nodes which are connected (not necessarily to be pairwisely connected) form a group. At the end, a vector of group labels can be obtained. The length of this vector is the same as the number of rows in the data matrix. Different rows with the same group label are in the same group. The number of distinct values in this label-vector is the number of groups.

**Value**

A vector of group labels, of the same length as the number of rows in the data matrix.

**References**


**Examples**

```r
nr=20
cmp=10
x=matrix(rnorm(nr*nc),nrow=nr,ncol=nc)
pearson.group(x,0.5)
```
skernel

Stratified Kernel

Description

This function defines a stratified kernel for metabolite abundance level measurements.

Usage

skernel(x, y, rho)

Arguments

x a numerical scalar or vector of metabolomic measurements.
y a numerical scalar or vector of metabolomic measurements.
rho a positive kernel shape parameter.

Details

This function calculates a stratified kernel function skernel between two metabolomic measurements x and y. Suppose the metabolite-set contains p metabolites. Then measurements x and y have p components. Let $x_i$ be the $i$th component of x. If $x_i = 0$, then the $i$th metabolite in the metabolite-set is absent. If $x_i > 0$, then the $i$th metabolite is present and $x_i$ measures the abundance level of the $i$th metabolite. Measurements x and y are said to from the same stratum if they have the same set of metabolites being absent (present). If x and y are from the same stratum, then $skernel(x, y, \rho)$ is assigned a Gaussian kernel with kernel parameter $\rho$. Otherwise $skernel(x, y, \rho)$ is defined to be 0. More details can be found in Zhan et al. (2015).

Value

A non-negative real value.

References


Examples

x=c(0,0,1,2)
y=c(0,1,2,0)
z=c(0,0,3,4)
## x and z are from the same stratum while x and y are not.
skernexp(x,y,1)
skernexp(x,z,1)
**spearman.group**  
*Grouping Based on Spearman Correlation Coefficients*

**Description**  
This function forms metabolite-sets based on pairwise Spearman correlation between different metabolites.

**Usage**  
spearman.group(data, threshold)

**Arguments**  
- **data**: a matrix with each row being a metabolite and each column being a sample.  
- **threshold**: a threshold value for correlation coefficients.

**Details**  
The input data is a matrix with each row denoting a metabolite. This function groups different rows of the data matrix together based on the Spearman correlation coefficients between two rows. It works in the following way.  
First, each row in the data matrix is treated as a node. If the Spearman correlation coefficient between two nodes is larger than the threshold value, then an edge is added between this two nodes. Second, all nodes which are connected (not necessarily to be pairwisely connected) form a group. At the end, a vector of group labels can be obtained. The length of this vector is the same as the number of rows in the data matrix. Different rows with the same group label are in the same group. The number of distinct values in this label-vector is the number of groups.

**Value**  
A vector of group labels, of the same length as the number of rows in the data matrix.

**References**  

**Examples**  
```r  
nr=20  
nc=10  
temp= sample(c(0,1,2,3), size=nr*nc, replace = TRUE, prob=c(0.4,0.2,0.2,0.2))  
x=matrix(temp,nrow=nr,ncol=nc)  
spearman.group(x,0.5)  ```
**Description**

This function test whether a metabolite-set is differential expressed using a stratified kernel-based score test.

**Usage**

```r
sscore(x, y, lower, upper, m)
```

**Arguments**

- `x`: numeric measurements of metabolite abundance level.
- `y`: 0/1 response indicating whether a subject is a case group or a control group.
- `lower`: lower bound of the kernel parameter.
- `upper`: upper bound of the kernel parameter.
- `m`: number of grid points selected in the interval \([lower, upper]\).

**Details**

Let \( x \) be a \( p \times n \) matrix, where each column is a subject, \( y \) be a \( n \times 1 \) 0/1 vector indicating the group label. This function tests whether this \( p \)-metabolite set is differentially expressed between two groups (more details can be found in Zhan et al. (2015)). It works in the following way.

A score test can be applied when the kernel parameter \( \rho \) is known. First, fit the null logistic model \( \logit(p(y = 1)) = \beta_0 \) to get estimate of \( \beta_0 \) as \( \hat{\beta}_0 \). Let \( \hat{\mu}_0 = \text{invlogit}(\hat{\beta}_0) \). Second, the kernel matrix is calculated as \( K(\rho)_{ij} = k(x_i, x_j, \rho) \), where \( x_i \) is \( i \)th column in \( x \), \( k(\cdot) \) is the stratified kernel function skernel. Third, the test statistic \( Q(\rho) \) is calculated as

\[
Q(\rho) = (y - \hat{\mu}_0)^T K(\rho) (y - \hat{\mu}_0).
\]

An standardized version \( S(\rho) \) of \( Q(\rho) \) can be calculated as \( S(\rho) = |Q(\rho) - \mu_Q|/\sigma_Q \). More details can be found in Liu et al. (2008).

When the kernel parameter \( \rho \) is not known. Suppose it takes values in \([lower, upper]\). Davies (1977) and Davies (1987) proposed a test based on the process \( \{S(\rho), \rho \in [lower, upper]\} \). This test has rejection region of the form \( \{\sup_{l \leq \rho \leq U} S(\rho) > c\} \). Using this test, an upper-bound for the \( p \)-value is given by:

\[
\Phi(-M) + V \exp(\frac{1}{2} M^2) / \sqrt{8 \pi},
\]

where \( \Phi(\cdot) \) is the cumulative distribution function of standard normal density, \( M \) is the maximum of \( S(\rho) \) over the range of \( \rho \) and \( V = [S(\rho_1) - S(lower)] + [S(\rho_1) - S(\rho_2)] + \cdots + [S(upper) - S(\rho_m)] \) is the total variation of \( S(\rho) \) over the interval \([lower, upper]\) and \( \rho_1, \ldots, \rho_m \) are \( m \) grid points in the interval \([lower, upper]\).
**Value**

A p-value indicating whether the metabolite-set is differentially expressed or not.

**References**

Davies, R. B. (1977) Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika, 64, 247-254.

Davies, R. B. (1987) Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika, 74, 33-43.


**See Also**

invlogit, skernel

**Examples**

data(hcc)

```r
x = hcc[1:3, 3:5]  # This metabolite-set contains the first three metabolites in the hcc dataset.
y = c(rep(0, 35), rep(1, 20))
sscore(x, y, 10^(-3), 10^3, 10)
```

---

**Description**

This function calculates the trace of a given numeric square matrix. This function should not be called directly in this package. It is called by other functions like dscore and sscore.

**Usage**

```r
tr(X)
```

**Arguments**

- **X** A square matrix

**Value**

A numeric value which is the sum of the values on the diagonal.
See Also

dscore, sscore

Examples

A <- matrix(seq(1:9), nrow=3, ncol=3)
tr(A)

tr
Index

dkernel, 2, 5
dscore, 3, 6, 13

hcc, 5

invlogit, 5, 6, 12

KMDA (KMDA-package), 2
KMDA-package, 2

mdist, 3, 7

pearson.group, 8

skernel, 9, 12
spearman.group, 10
sscore, 6, 11, 13

tr, 12